Uniform Crime Reporting (UCR) Program Data: A Practitioner's Guide

Jacob Kaplan

2021-03-20

Contents

Preface

If you've read an article about crime or arrests in the United States in the last half century, in most cases it was referring to the FBI's Uniform Crime Reporting Program Data, otherwise known as UCR data. UCR data is, with the exception of the more detailed data that only covers murders, a *monthly number of crimes or arrests reported to a single police agency* which is then gathered by the FBI into one file that includes all reporting agencies. Think of your home town. This data will tell you how many crimes were reported for a small number of crimes or how many people (broken down by age, sex, and race) were arrested for a (larger) set of crimes in that city (if the city has multiple police agencies, it will use the agency which is the primary agency on the case, usually the local police department) in a given month. This is a very broad measure of crime, and its uses in research - or understanding crime at all - is fairly limited. Yet is has become - and will likely remain among researchers for at least the next decade - the most important crime data in the United States.

UCR data is important for three reasons:

1. The definitions are standard and all agencies follow them so you can compare across agencies and over time
2. The data is available since 1960 so there is a long period of available data
3. The data is available for most of the 18,000 police agencies in the United States so you can compare across agencies

For most of this book we'll be discussing the caveats of the above reasons - or, more directly, why these assumptions are wrong - but these are the reasons why the data is so influential.

Motivation

By the end of each chapter you should have a firm grasp on the dataset the covered and how to use it properly. However, this book can't possibly cover every potential use case for the data so make sure to carefully examine the data for your own particular use. This benefits you because you'll know your data better and become a better research because of it. This benefits me because it'll increase the quality of research in my field.

I get a lot of emails from people asking questions about this data so part of my motivation for writing this book is to create a single place that answers as many questions as I can about the data. Again, this is among the most commonly used crime datasets and there are still many current papers published with incorrect information about the data (including such simple aspects like what geographic unit data is in and what time unit it is in). So hopefully this book will decrease the amount of misconceptions about this data, increasing overall research quality.[1]

Structure of the book

This remainder of this book will be divided into eight chapters: an intro chapter briefly summarizing each dataset and going over overall issues with UCR data, six chapters each covering one of the six UCR datasets, and a final one covering county-level data, a highly flawed but common use of the UCR data. Each chapter will follow the same format: we'll start with a history of the data such as when it first because available and important changes in definitions or variables.

Next we'll discuss what the data looks like initially when you get it from the FBI - literally what it looks like in it's fixed-width ASCII format you get from the FBI and what it looks like (what each row and column mean) once it's turned into a useful format that can be read into modern software like R and Stata.[2] For most of the datasets this is a minor process but for data like

[1] Ideally, this will also decrease the number of emails and increase the number of citations I receive.

[2] To look at the data in its fixed-width ASCII format I'll use the program Notepad which can open up text files like ASCII files. For looking at the machine-readable format

the arrest or homicide datasets, the conversion process is harder - and this can actually lead to changes in the resulting data. For example, in an old version of the arrest data that I released, I aggregated certain arrestee ages together since my laptop at the time couldn't handle converting data from ASCII to R and Stata without aggregating the ages (more age groups means more columns which means more computer memory needed). So anyone using that data would have less detailed data than the current dataset.

Understanding how the data moves from its rawest form (which in this case is after cleaning by the FBI) is important for being able to truly understand the data and its caveats. However, this is a fairly technical part of each chapter so feel free to skip it. Next in each chapter, we'll cover the variables included in the data and how to use them properly (including not using them at all) - this will be the bulk of each chapter. We'll end each chapter by briefly summarizing the data, how and when it's useful, and - most importantly - when you shouldn't use it.

Since manuals are boring, I'll try to include graphs and images to try to alleviate the boredom. That said, I don't think it's possible to make it too fun so sorry in advanced. This book is a mix of facts about the data, such as how many years are available, and my opinions about it, such as whether it is reliable. In cases of facts I'll just say a statement - e.g. "the offenses data is available since 1960". In cases of opinion I'll temper the statement by saying something like "in my opinion…"

Citing this book

If this book was useful in your research, please cite it. To cite this book, please use the below citation:

Kaplan J (2021). *Uniform Crime Reporting (UCR) Program Data: A Practitioner's Guide*. https://github.com/jacobkap/ucrbook.

BibTeX format:

I'll use R since I think it's a bit better looking than viewing it in Stata. In both cases, these will be images included in the chapter - you won't need to follow along or use either program.

Sources of UCR data

There are a few different sources of UCR data available today. First, and probably most commonly used, is the data put together by the National Archive of Criminal Justice Data (NACJD)). This a team of out of the University of Michigan who manages a huge number of criminal justice datasets and makes them available to the public. If you have any questions about crime data - UCR or other data - I highly recommend you reach out to them for answers. They have a collection of data and excellent documentation available for UCR data available on their site here. One limitation to their data, however, is that each year of data is available as an individual file meaning that you'll need to concatenate each year together into a single file. Some years also have different column names (generally minor changes like spelling robbery "rob" one year and "robb" the next) which requires more work to standardize before you could concatenate. They also only have data through 2016 which means that the most recent years (UCR data is available through 2019) of data are (as of this writing) unavailable.

Next, and most usable for the general public but limited to researchers, is the FBI's official website Crime Data Explorer. On this site you can chose an agency and see annual crime data (remember, UCR data is monthly so this isn't as detailed as it can be) for certain crimes. This is okay for the general public but only provides a fraction of the data available in the actual data so is really not good for researchers.

Finally, I have my own collection of UCR data available publicly on openICPSR, a site which allows people to submit their data for public access. For each of these datasets I've taken the raw data from the FBI (for early years of homicide data this is actually from NACJD since the FBI's raw data is wrong and can't be parsed. For later years of homicide data this is from the FBI's raw data.) and read it into R. Since the data is only available from the FBI as fixed-width ASCII files, I created a setup file (we'll explain exactly how reading in this kind of data works in the next chapter) and read the data and then very lightly cleaned the data (i.e. only removing extreme outliers like an agency having millions of arsons in a month). For each of these datasets I detail what I've done to the data and briefly summarize the data (i.e. a very short version of this book) on the data's page on openICPSR. The main advantage is that all my data has standard variable names and column names and, for data that is small

enough, provide the data as a single file that has all years. For large datasets like the arrest data I break it down into parts of the data and not all years in a single file. The downside is that I don't provide documentation other than what's on the openICPSR page and only provide data in R and Stata format. I also have a similar site to the FBI's Crime Data Explorer but with more variables available, that site is available here.

It's worth mentioning a final source of UCR information. This is the annual Crimes in the United States report released by the FBI each year around the start of October.[3] As an example, here is the website for the 2019 report. In this report is summarized data which in most cases estimates missing data and provides information about national and subnational (though rarely city-level) crime data. As with the FBI's site it is only a fraction of the true data available so is not a very useful source of crime data. Still, this is a very common source of information used by researchers.

Where to find the data used in this book

The data I am using in this book is the cleaned (we'll discuss in more detail exactly what I did to clean each dataset in the dataset's chapter but the short answer is that I did very little.) and concatenated data that I put together from the raw data that the FBI releases. That data is available on my website here. I am hosting this book through GitHub which has a maximum file size allowed that is far smaller than these data so you'll need to go to my site to download the data, it's not available through this book's GitHub repo. For some examples I'm using the data before I cleaned it of outliers (as an example of the outliers present before I removed them) so that data is not publicly available.

NIBRS data

Another source of FBI data, and one sometimes considered part of the UCR data collection, is the National Incident-Based Reporting System (NIBRS)

[3] They also release a report about the first 6-months of the most recent year of data before the October release but this is generally an estimate from a sample of agencies so is far less useful.

data. Like its name implies this is an incident-level dataset which has detailed information about each incident reported to the police, including incident circumstances, and victim and offender information. This is also the data that the FBI has declared will replace UCR data starting in 2021, meaning that they will no longer collect UCR data and only allow agencies to submit NIBRS data. NIBRS data is a complex and highly rich dataset that deserves its own book to really understand, so I will not be discussing it any further in this book.

About the author

Jacob Kaplan holds a PhD and a master's degree in criminology from the University of Pennsylvania and a bachelor's degree in criminal justice from California State University, Sacramento. His research focuses on Crime Prevention Through Environmental Design (CPTED), specifically on the effect of outdoor lighting on crime. He is the author of several R packages that make it easier to work with data, including fastDummies and asciiSetupReader. His website allows easy analysis of crime-related data and he has released over a dozen crime data sets (primarily FBI UCR data) on openICPSR that he has compiled, cleaned, and made available to the public.

For a list of papers he has written (including working papers), please see here.

For a list of data sets he has cleaned, aggregated, and made public, please see here.

For a list of R packages he has created, please see here.

Chapter 1

Overview of the Data

One of the first, and most important questions, I think people have about crime is a simple one: is crime going up? Answering it seems simple - you just count up all the crimes that happen in an area and see if that number of bigger than it was in previous times.

However, putting this into practice invites a number of questions, all of which affect how we measure crime. First, we need to define what a crime is? Not philosophically what actions are crimes - or what should be crimes - but literally which of the many thousands of different criminal acts (crimes as defined by state law) should be considered in this measure. Should murder count? Most people would say yes. How about jaywalking or speeding? Many would say probably not. Should marital rape be considered a crime? Now, certainly most people (all, I would hope) would say yes. But in much of the United States it wasn't a crime until the 1970s (Bennice and Resick, 2003; McMahon-Howard et al., 2009).

Next, we have to know what geographic and time unit to measure crimes at since these decisions determine how precise we can measure crime and when it changed. If we

The final question is that when a crime occurs, how do we know? That is, when we want to count how many crimes occurred do we ask people how often they've been victimized, do we ask people how often they commit a crime, do we look at crimes reported to police, crimes charged in a criminal court? Each of these measures will likely give different answers as to how

many crimes occurred.[1]

The FBI answered all of these questions in 1929 when they began the Uniform Crime Reporting (UCR) Program Data, or UCR data for short. Crime consists of seven crime categories - murder, rape, robbery, aggravated assault, burglary, motor vehicle theft, and theft - that are reported to the police and is collected each month from each agency in the country. These decisions, born primarily out of the resource limitations of 1929 (e.g. no computers), have had a major impact on criminology research. These seven crime categories - known as "Index Crimes" or "Part 1 crimes" (or "Part I" sometimes) - are the ones used to measure crime in many criminology papers, even when the researchers have access to data that covers a broader selection of crimes than these.[2]

1.0.1 The data as you get it from the FBI

We'll finish this overview of the UCR data by briefly talking about format of the data that is released by the FBI, before the processing done by myself or NACJD that converts the data to a type that software like R or Stata or Excel can understand. The FBI releases their data as fixed-width ASCII files which are basically just an Excel file but with all of the columns squished together. As an example, below is the data as you receive it from the FBI for the Offenses Known and Clearances by Arrest dataset for 1960, the first year with data available. In the figure, it seems like there are multiple rows but that's just because the software that I opened the file in isn't wide enough - in reality what is shown is a single row that is extremely wide because there are over 1,500 columns in this data. If you scroll down enough you'll see the next row, but that isn't shown in the current image. What is shown is a single row with a ton of columns all pushed up next to each other. Since all of the columns are squished together (the gaps are just blank spaces because the value there is a space, but that doesn't mean there is a in the data.

[1] The Bureau of Justice Statistics does measure crime by asking a random sample of people whether they were the victim of a crime. For more on this, please see their National Crime Victimization Survey reports

[2] Arson is also an index crime but was added after these initial seven were chosen and is not included in the crimes dataset so is generally not included in studies that use index crimes.

Spaces are possible values in the data and are meaningful), you need some way to figure out which parts of the data belong in which column.

```
knitr::include_graphics('images/offenses_known_raw_ascii_1960.PNG')
```

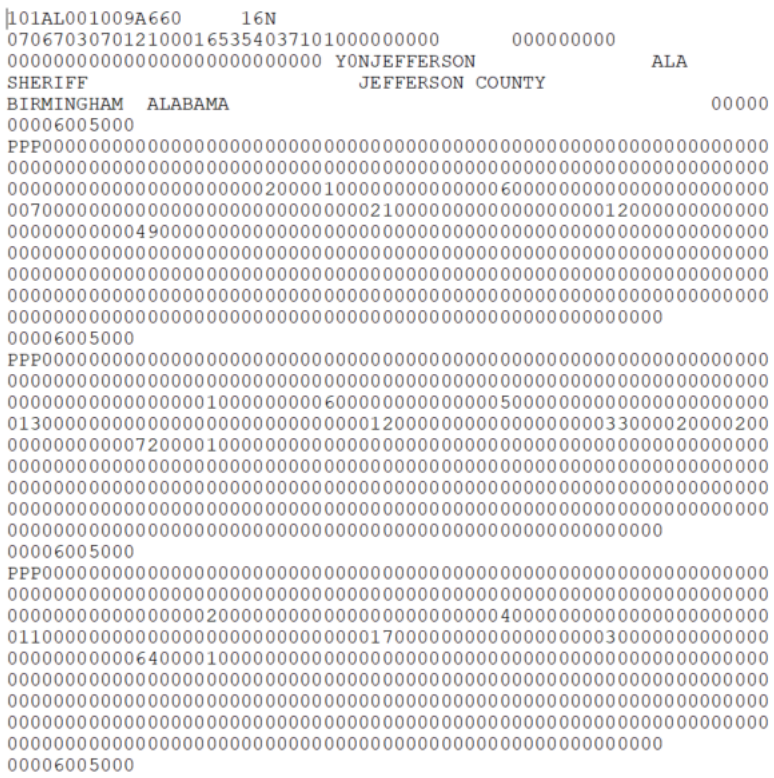

Figure 1.1: Fixed-width ASCII file for the 1960 Offenses Known and Clearances by Arrest dataset

The "fixed-width" part of the file type is how this works (the ASCII part basically means it's a text file). Each row is the same width - literally the same number of characters, including blank spaces. So you must tell the software you are using to process this file - by literally write code in something called a "setup file" but is basically just instructions for whatever software you use (R, SPSS, Stata, SAS can all do this) - which characters are certain columns. For example, in this data the first character says which type of UCR data it is (1 means the Offenses Known and Clearances by Arrest

data) and the next two characters (in the setup file written as 2-3 since it is characters 2 through 3 [inclusive]) are the state number (01 is the state code for Alabama). So we can read this row as the first column indicating it is an Offenses Known data, the second column indicating that it is for the state of Alabama, and so on for each of the remaining columns. To read in this data you'll need a setup file that covers every column in the data (some software, like R, can handle just reading in the specific columns you want and don't need to include every column in the setup file).

The second important thing to know about reading in a fixed-width ASCII file is something called a "value label."[3] For example, in the above image we saw the characters 2-3 is the state and in the row we have the value "01" which means that the state is "Alabama." Since this type of data is trying to be as small as efficient as possible, it often replaces longer values with shorter one and provides a translation for the software to use to convert it to the proper value when reading it. "Alabama" is more characters than "01" so it saves space to say "01" and just replace that with "Alabama" later on. So "01" would be the 'value" and "Alabama" would be the 'label' that it changes to once read.

Fixed-width ASCII files may seem awful to you reading it today, and it is awful to use. But it appears to be an efficient way to store data back many decades ago when data releases began but now is extremely inefficient - in terms of speed, file size, ease of use - compared to modern software so I'm not sure why they *still* release data in this format. But they do, and even the more *modern* (if starting in 1991, before I was born, is modern!) NIBRS data comes in this format. For you, however, the important part to understand is not how exactly to read this type of data, but to understand that people who made this data publicly available (such as myself and the team at NACJD) must made this conversion process.[4] **This conversion process, from fixed-width ASCII to a useful format is the most dangerous step taken in using this data - and one that is nearly entirely unseen by researchers.**

[3]For most fixed-width ASCII files there are also missing values where it'll have placeholder value such as -8 and the setup file will instruct the software to convert that to NA. UCR data, however, does not have this and does not indicate when values are missing in this manner.

[4]For those interested in reading in this type of data, please see my R package asciiSetupReader.

Every line of code you write (or, for SPSS users, click you make) invites the possibility of making a mistake.[5] The FBI does not provide a setup file with the fixed-width ASCII data so to read in this data you need to make it yourself. Since some UCR data are massive, this involves assigning the column width for thousands of columns and the value labels for hundreds of different value labels.[6] A typo anywhere could have potentially far-reaching consequences so this is a crucial weakpoint in the data cleaning process - and one in which I have not seen anything written about before.

1.1 What is a crime?

1.1.1 Crimes included in the UCR datasets

1.1.2 Index crimes

1.1.3 Violent crimes

1.2 A summary of each UCR dataset

UCR data can be roughly summarized into two groups: crime data and arrest data. While there are several datasets included they are all extensions of one of the above groups. For arrest data, you have information about who (by race and by age-gender, but not by race-gender or race-age other than within race you know if the arrestee is an adult or a juvenile). For crime data, you have monthly counts of a small number of crimes (many fewer than crimes covered in the arrest data) and then more specialized data on a subset of these crimes - information on homicides, hate crimes, and property stolen.

[5] Even highly experienced programmers who are doing something like can make mistakes. For example, if you type out "2+2" 100 times - something extremely simple that anyone can do - how often will you mistype a character and get a wrong result? I'd guess that at least once you'd make a mistake.

[6] With the exception of the arrest data and some value label changes in hate crimes and homicide data, the setup files remain consistent you a single file will work for all years for a given dataset. You do not need to make a setup file for each year.

1.2.1 Offenses Known and Clearances by Arrest

This dataset is the oldest and most widely used of the UCR datasets. It covers

1.2.2 Arrests by Age, Sex, and Race

1.2.3 Law Enforcement Officers Killed and Assaulted (LEOKA)

1.2.4 Supplementary Homicide Reports (SHR)

1.2.5 Hate Crime Data

1.2.6 Property Stolen and Recovered (Supplement to Return A)

Chapter 2

Offenses Known and Clearances by Arrest

2.1 A brief history of the data

2.1.1 Changes in definitions

2.2 What does the data look like?

2.3 What variables are in the data?

2.3.1 Key variables

2.3.2 Known issues with the data

2.4 Final thoughts

2.5 ORIs - Unique agency identifiers

In the UCR and other FBI data sets, agencies are identified using **OR**iginating Agency **I**dentifiers or ORIs. These are unique ID codes used

to identify an agency. If we used the agency's name we'd end up with some duplicates For example, if you looked for the Philadelphia Police Department using the agency name, you'd find both the "Philadelphia Police Department" in Pennsylvania and the one in Mississippi.

Each ORI is a 7-digit value starting with the state abbreviation (for some reason the FBI incorrectly puts the abbreviation for Nebraska as NB instead of NE) followed by 5 numbers. In the NIBRS data (another FBI data set) the ORI uses a 9-digit code - expanding the 5 numbers to 7 numbers. When dealing with specific agencies, make sure to use the ORI rather than the agency name to avoid any mistakes.

For an easy way to find the ORI number of an agency, use this site. Type an agency name or an ORI code into the search section and it will return everything that is a match.

2.6 Hierarchy Rule

This dataset uses what is called the Hierarchy Rule where only the most serious crime in an incident is reported (except for motor vehicle theft, which is always included). For example if there is an incident where the victim is robbed and then murdered, only the murder is counted as it is considered more serious than the robbery.

How much does this affect our data in practice? Actually relatively little. Though the Hierarchy Rule does mean this data is an under-count, data from other sources indicate that it isn't much of an under count. The FBI's other data set, the National Incident-Based Reporting System (NIBRS) contains every crime that occurs in an incident (i.e. it doesn't use the Hierarchy Rule). Using this we can measure how many crimes the Hierarchy Rule excludes (Most major cities do not report to NIBRS so what we find in NIBRS may not apply to them). In over 90% of incidents, only one crime is committed. Additionally, when people talk about "crime" they usually mean murder which, while incomplete to discuss crime, means the UCR data here is accurate on that measure.

2.7 Which crimes are included?

If you look back at the output when we ran `names(offenses_known_yearly_1960_2017)` you'll see that it produced five broad categories of columns. The first was information about the agency including population and geographic info, then came four columns with the same values except starting with "actual", "tot_clr", "clr_18", and "unfound". Following these starting values were 30 crime categories. We'll discuss what each of those starting values mean in a bit, let's first talk about which crimes are included and what that means for research.

2.7.1 Index Crimes

The Offenses Known and Clearances by Arrest data set contains information on the number of "Index Crimes" (sometimes called Part I crimes) reported to each agency. These index crimes are a collection of eight crimes that, for historical reasons based largely by perceived importance in the 1920's when the UCR program was first developed, are used as the primary measure of crime today. Other data sets in the UCR, such as the Arrests by Age, Sex, and Race data and the Hate Crime data have more crimes reported.

The crimes are, in order by the Hierarchy Rule -

1. Homicide
    - Murder and non-negligent manslaughter
    - Manslaughter by negligence
2. Rape
    - Rape
    - Attempted rape
3. Robbery

- With a firearm
- With a knife of cutting instrument
- With a dangerous weapon not otherwise specified
- Unarmed - using hands, fists, feet, etc.

4. Aggravated Assault (assault with a weapon or causing serious bodily injury)
   - With a firearm
   - With a knife of cutting instrument
   - With a dangerous weapon not otherwise specified
   - Unarmed - using hands, fists, feet, etc.

5. Burglary
   - With forcible entry
   - Without forcible entry
   - Attempted burglary with forcible entry

6. Theft (other than of a motor vehicle)

7. Motor Vehicle Theft
   - Cars
   - Trucks and buses
   - Other vehicles

8. Arson

9. Simple Assault

For a full definition of each of the index crimes see the FBI's Offense Definitions page here.

Arson is considered an index crime but is not reported in this data - you need to use the separate Arson data set of the UCR to get access to arson counts. The ninth crime on that list, simple assault, is not considered an index crime but is nevertheless included in this data.

Each of the crimes in the list above, and their subcategories, are included in the UCR data. In most reports, however, you'll see them reported as the total number of index crimes, summing up categories 1-7 and reporting that as "crime." These index crimes are often divided into violent index crimes - murder, rape, robbery, and aggravated assault - and property index crimes - burglary, theft, motor vehicle theft.

2.7.2 The problem with using index crimes

The biggest problem with index crimes is that it is simply the sum of 8 (or 7 since arson data usually isn't available) crimes. Index crimes have a huge range in their seriousness - it includes both murder and theft. This is clearly wrong as 100 murders is more serious than 100 thefts. This is especially a problem as less serious crimes (theft mostly) are far more common than more serious crimes (in 2017 there were 1.25 million violent index crimes in the United States. That same year had 5.5 million thefts.). So index crimes under-count the seriousness of crimes. Looking at total index crimes is, in effect, mostly just looking at theft.

This is especially a problem because it hides trends in violent crimes. San Francisco, as an example, has had a huge increase in index crimes in the last several years. When looking closer, that increase is driven almost entirely by the near doubling of theft since 2011. During the same years, violent crime has stayed fairly steady. So the city isn't getting more dangerous but it appears like it is due to just looking at total index crimes.

Many researchers divide index crimes into violent and nonviolent categories, which helps but is still not entirely sufficient. Take Chicago as an example.

It is a city infamous for its large number of murders. But as a fraction of index crimes, Chicago has a rounding error worth of murders. Their 653 murders in 2017 is only 0.5% of total index crimes. For violent index crimes, murder makes up 2.2%. What this means is that changes in murder are very difficult to detect. If Chicago had no murders this year, but a less serious crime (such as theft) increased slightly, we couldn't tell from looking at the number of index crimes.

2.7.3 Rape definition change

Starting in 2013, rape has a new, broader definition in the UCR to include oral and anal penetration (by a body part or object) and to allow men to be victims. The previous definition included only forcible intercourse against a woman. As this revised definition is broader than the original one post-2013, rape data is not comparable to pre-2013 data.

2.8 Actual offenses, clearances, and unfounded offenses

For each crime we have four different categories indicating the number of crimes actually committed, the number cleared, and the number determined to not have occurred.

2.8.1 Actual

This is the number of offenses that occurred, simply a count of the number of crimes that month. For example if 10 people are murdered in a city the number of "actual murders" would be 10.

2.8.2 Total Cleared

A crime is cleared when an offender is arrested or when the case is considered cleared by exceptional means. When a single offender for a crime is arrested, that crime is considered cleared. If multiple people committed a crime, only

a single person must be arrested for it to be cleared, and as the UCR data is at the offense level, making multiple arrests for an incident only counts as one incident cleared. So if 10 people committed a murder and all 10 were arrested, it would report one murder cleared not 10. If only one of these people are arrested it would still report one murder cleared - the UCR does not even say how many people commit a crime.

A crime is considered exceptionally cleared if the police can identify the offender, have enough evidence to arrest the offender, know where the offender is, but is unable to arrest them. Some examples of this are the death of the offender or when the victim refuses to cooperate in the case.

Unfortunately this data does not differentiate between clearances by arrest or by exceptional means. For a comprehensive report on how this variable can be exploited to exaggerate clearance rates, see this report by ProPublica on exceptional clearances with rape cases.

2.8.3 Cleared Where All Offenders Are Under 18

This variable is very similar to Total Cleared except is only for offenses in which **every** offender is younger than age 18.

2.8.4 Unfounded

An unfounded crime is one in which a police investigation has determined that the reported crime did not actually happen. For example if the police are called to a possible burglary but later find out that a burglary did not occur, they would put it down as 1 unfounded burglary. This is based on police investigation rather than the decision of any other party such as a coroner, judge, jury, or prosecutor.

2.9 Number of months reported

UCR data is reported monthly though even agencies that decide to report their data may not do so every month. As we don't want to compare an

agency which reports 12 months to one that reports fewer, the variable *number_of_months_reported* is way keep only agencies that report 12 months, or deal with those that report fewer.

From our `table()` output it seems that when agencies do report, they tend to do so for all 12 months of the year. However, this variable is seriously flawed, and its name is quite misleading. In reality this variable is actually just whichever the last month reported was. If an agency reported every month of the year, meaning December is the last month, they would have a value of 12. If the agency **only** reported in December, they would also have a value of 12. While there are ways in the monthly data to measure actual number of months reported, these ways are also flawed. So be cautious about this data and particularly the value of this variable.

Chapter 3

Arrests by Age, Sex, and Race

It differs from the clearance information in the Offenses Known and Clearances by Arrest data since this data includes a wider variety of crimes and breaks down arrests into more detailed bins.

3.1 A brief history of the data

3.1.1 Changes in definitions

3.2 What does the data look like?

3.2.1 Raw data

3.3 What variables are in the data?

3.3.1 Key variables

3.4 Known issues with the data

3.5 Final thoughts

Chapter 4

Law Enforcement Officers Killed and Assaulted (LEOKA)

The Law Enforcement Officers Killed and Assaulted data, often called just by its acronym LEOKA, has two main purposes.[1] First, it provides counts of employees employed by each agency - broken down by if they are civilian employees or sworn officers, and also broken down by gender. And second it measures how many officers were assaulted or killed (including officers who die accidentally such as in a car crash) in a given month - this is broken down into shift type and type (e.g. alone, with a partner, on foot, in a car, etc.), the offender's weapon, and type of call they are responding to (e.g. robbery, disturbance, traffic stop).

The employee information is at the year-level so you know, for example, how many male police officers were employed in a given year at an agency, but don't know any more than that such as how many officers were on patrol, were detectives, were in special units, etc. This dataset is commonly used as a measure of police employees and is a generally reliable - though imperfect as we'll see - measure of how many police are employed by a police agency. The second part of this data, measuring assaults and deaths, is more flawed with missing data issues and data error issues (e.g. more officers killed than employed in an agency). We'll get into these issues in more detail during this chapter.

[1] This data is also sometimes called the "Police Employees" dataset.

This data, as well as the privately-run site Officer Down Memorial Page which covers law enforcement officers who have died, has also been used lately in the context of police using force against people out of fear of being harmed by that person. The discussion revolves around whether police are actually in high danger of being harmed by comparing the rate at which officers die to that of other professions. In general they find that police officers are among the most likely profession to die but are not at the top of this measure.

4.1 A brief history of the data

4.1.1 Changes in definitions

4.2 What does the data look like?

4.3 What variables are in the data?

4.3.1 Key variables

4.4 Known issues with the data

4.5 Final thoughts

Chapter 5

Supplementary Homicide Reports (SHR)

This is the most detailed of the UCR datasets and provides information about the circumstances and participants (victim and offender demographics and relationship status) for homicides.[1] For each homicide incident it tells you the age, gender, race, and ethnicity of each victim and offender as well as the relationship between the first victim and each of the offenders (but not the other victims in cases where there are multiple victims). It also tells you the weapon used by each offender and the circumstance of the killing, such as a "lovers triangle" or a gang-related murder. As with other UCR data, it also tells you the agency it occurred in and the month and year when the crime happened.

While highly detailed compared to other UCR data, there are a number of limitations for this data.

Since this data is voluntary to

This
If this "most detailed" dataset sounds disappointing - and it is! -

[1] If you're familiar with the National Incident-Based Reporting System (NIBRS) data that is replacing UCR, this dataset is the closest UCR data to it, though it is still less detailed than NIBRS data.

5.1 A brief history of the data

The data is available from the FBI starting in 1975 though, unlike all later years, this year only has information on a single victim and a single offender. For this reason I only release data starting in 1976 where up to 11 victims and 11 offenders are included. This data has been released every year since and the most recent year available is 2019.

5.1.1 Changes in definitions

5.2 What does the data look like?

5.2.1 Raw data

5.3 What variables are in the data?

5.3.1 Key variables

5.4 Known issues with the data

5.5 Final thoughts

Chapter 6

Hate Crime Data

This dataset covers crimes that are reported to the police and judged by the police to be motivated by hate More specifically, they are, first, crimes which were, second, motivated - at least in part - by bias towards a certain person or group of people because of characteristics about them such as race, sexual orientation, or religion. The first part is key, they must be crimes - and really must be the selection of crimes that the FBI collects for this dataset. Biased actions that don't meet the standard of a crime, or are of a crime not included in this data, are not considered hate crimes. For example, if someone yells at a Black person and uses racial slurs against them, it is clearly a racist action. For it to be included in this data, however, it would have to extend to a threat since "intimidation" is a crime included in this data but lesser actions such as simply insulting someone is not included. For the second part, the bias motivation, it must be against a group that the FBI includes in this data. When this data collection began crimes against transgender people were not counted so if a transgender person was assaulted or killed because they were transgender, this is not a hate crime recorded in the data.[1]

So this data is really a more narrow measure of hate crimes than it might seem. In practice it is (some) crimes motivated by (some) kinds of hate that are reported to the police. It is also the most under-reported UCR dataset with most agencies not reporting any crimes to the FBI. This leads to huge gaps in the data with some states having zero agencies report crime,

[1]The first year where transgender as a group was a considered a bias motivation was in 2014.

agencies reporting some bias motivations but not others, agencies reporting some years but not others. While these problems exist for all of the UCR datasets, it is most severe in this data. This problem is exacerbated by hate crimes being rare even in agencies that report them - with such rare events, even minor changes in which agencies report or which types of offenses they include can have large effects.

My main takeaway for this data is that it is inappropriate to use it to study hate crimes. At most it can be used to look at within-city within-bias-motivation trends, while keeping in mind that even this narrow subset of data is limited by under-reporting by victims and potential changes in police practices of reporting such as how many months of data they report per year.

6.1 A brief history of the data

6.1.1 Changes in definitions

6.2 What does the data look like?

6.2.1 Raw data

6.3 What variables are in the data?

6.3.1 Key variables

6.4 Known issues with the data

6.5 Final thoughts

Chapter 7

Property Stolen and Recovered (Supplement to Return A)

7.1 A brief history of the data

7.1.1 Changes in definitions

7.2 What does the data look like?

7.2.1 Raw data

7.3 What variables are in the data?

7.3.1 Key variables

7.4 Known issues with the data

7.5 Final thoughts

Chapter 8

County-Level Detailed Arrest and Offense Data

8.1 A brief history of the data

8.1.1 Changes in definitions

8.2 What does the data look like?

8.2.1 Raw data

8.2.2 Cleaned data

8.3 What variables are in the data?

8.3.1 Key variables

8.3.2 Known issues with the data

8.4 Final thoughts

Bibliography

Bennice, J. A. and Resick, P. A. (2003). Marital rape: History, research, and practice. *Trauma, Violence, & Abuse*, 4(3):228–246.

McMahon-Howard, J., Clay-Warner, J., and Renzulli, L. (2009). Criminalizing spousal rape: The diffusion of legal reforms. *Sociological Perspectives*, 52(4):505–531.